Visually Exploring Data Provenance and Quality of Open Data

C. Bors[1], T. Gschwandtner[1], and S. Miksch[1]

[1]TU Wien, Institute of Visual Computing and Human-Centered Technology, Austria

Abstract

While open data platforms are increasingly popular among end-users as well as data providers, there is a growing problem with inconsistent update frequencies and lack of quality in datasets. Efforts to monitor data quality are currently limited to checking meta-information and creating revisions to allow manual inspection of former datasets. We employ a Visual Analytics framework for generating and visualizing data provenance from data quality to facilitate data analysis and help users to understand the impact of updates on the data. Data quality metrics are utilized to quantify the development of data quality over time for open data projects. We combine quality metrics, data provenance, and data transformation information in an interactive exploration environment to expedite assessment and selection of appropriate open datasets.

CCS Concepts

•***Information systems*** → ***Data cleaning;*** *Data analytics;* •***Human-centered computing*** → ***Visual analytics;***

1. Introduction

In recent years open datasets are becoming readily available on diverse platforms, being provided by organizations, governments, and even companies, promoting the use of data for miscellaneous purposes. There is a high variability in the type and source of data, and thus also a high variance in quality of available data and their update frequency, which depends on the ability and willingness of providers to correct existing data or submit new valid data. Users of the data are often left with a dataset of poor quality without context or insights into what problems persist in the data and how they can be resolved. In this poster we present an approach to generating data provenance from data quality aspects to make the data processing pipeline more comprehensible and give users understanding in how quality was affected by the individual processes. By retaining provenance during the iterative processes of data cleansing [GGAM12, GAM*14] and wrangling [KHP*11] we can observe varying assessment strategies and quantitatively compare different results by calculating data quality metrics throughout analysis. Data quality metrics, measures of data quality dimensions that give proportional information of the lack of quality with respect to a particular context or aspect [Sad13], make data properties quantifiable so that they can be subsequently communicated visually.

Interactive data quality analysis is comprised of the tasks of data profiling [KPP*12], wrangling and cleansing with the goal to transform the data into usable form for subsequent analysis. Data profiling approaches assist users throughout the quality assessment process by enabling effective exploration of different aspects of the data. However, after assessment these exploration abilities are lost along with the transformation history. No retrospective perspective analysis view is available to users, which means users have to constantly re-use cleansing, profiling, or wrangling environments to validate data. But for datasets that have been edited and processed by other users – which is increasingly likely due to portals offering multiple revisions of datasets – it is not comprehensible to revisit the applied processing steps. Capturing data provenance [RESC16] is used to store various workflow processes, comprehending visualization workflows, or contextualize development of data over time. However, data quality aspects are often not communicated to users. This information could aid users in assessing the current state of an open dataset and allow them to observe the development of the data over time and check the update frequency. That way they can validate if data quality is sufficient for their analysis purposes. Including data transformations and data quality assessment aspects as data provenance generates traceable data, in combination with a history of changes over time which other users can make use of them to expedite data analysis.

2. Generating Quality Provenance of Open Data

The *ADEQUATe* project [ade17] provides a community driven open data service that periodically assesses open data sets for changes and quality and leverages crowdsourcing to identify and resolve quality issues in open data projects. However, qualitative aspects are only derived from meta-information, with users still required to manually check the list of changes and inspect raw data for validation. We propose a Visual Analytics (VA) approach to integrate data quality into open data provenance. To accomplish this, we developed a framework that derives data provenance from data quality analysis and processing operations and incorporated a data

DOI: 10.2312/eurp.20181117

wrangling tool into the open data service that enables users to draw information from exploring changes of data quality throughout data revisions over time, observing the impact of data transformations on quality and data structure, branching data to qualitatively compare different approaches. Our framework captures provenance from changes to the data and retains the data's history through snapshots and versioning control. Additionally, the provenance information is extended with data quality metrics that give contextual information on changes of quality [BKG*18]. For that we store each data revision in an existing server back-end (OpenRefine [MGM12] server) that features quality assessment and data wrangling functionality. Combined with the derived provenance, we can trace which changes had what kind of impact on data quality along with provenance generated from logging data revisions and transformations. Insights can be generated by exploring the raw data or a list of changes in the form of a `git diff`† file. This allows to leverage information about what changes were done to the data along with insights into how they affected data quality and, subsequently, if the data are usable for analysis.

Quantifying Data Quality Our VA framework is fitted with a metrics recommendation system for tabular data – based on data properties (e.g., data types, distribution information) – which computes type-appropriate metrics [BKG*18], like a general completeness metric, a statistical plausibility measure for numeric values, or a validity metric for strings, numerical, or text values. These can be adapted by users to accommodate for domain specific characteristics [BKG*18]. By logging data quality after any revision, we can quantify changes in quality over time, enabling comparison between datasets, different revisions, as well as different processing pipelines, if available. By computing the qualitative changes over time, along with meta-information of the data and the corresponding source of change, users are able to judge whether this change had a positive or negative impact.

2.1. A Visual Analytics Approach

Our VA framework features multiple dedicated panels, with the main view being comprised of a composite visualization (see Fig. 1), combining provenance information – gathered alongside data transformation operations and data revisions – as a provenance graph, consolidated data quality information visualized over time via stacked bar charts, and meta-information enhancing the provenance graph with contextual information (e.g., Fig. 1a, bars for additions and deletions per commit). The quality information indicates detected quality issues proportional to the entire dataset, for each column respectively: If the column of a dataset contains a large number of quality problems w.r.t. a particular metric, the stacked bar chart is large for this column and metric (see Fig. 1c). This composite visualization is used to iteratively explore the project data and navigate into dedicated exploration views, where users can explore data quality, the raw data, or the history of changes individually. The user is encouraged to explore different states of the data to understand and comprehend changes to the data. The framework also allows customization of quality metrics,

† https://git-scm.com/docs/git-diff

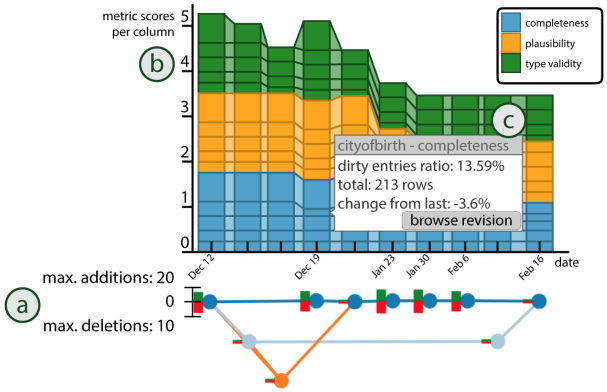

Figure 1: *The composite data quality provenance view combines (a) an enhanced provenance graph of revisions and transformations, with (b) a stacked barchart visualization of quality metrics to allow qualitative comparison between versions and (c) get details on demand for particular individual metrics per column. The graph coloring corresponds to individual branches in the project.*

these are also stored as provenance, hence changes in the dataset can be accounted for and quality information remains meaningful and can be useful to others. With our framework we aim to externalize an facilitate the tasks of **(a)** analyzing the development and evolution of a dataset, **(b)** exploring multiple revisions of a dataset, **(c)** getting detailed information on quality information to judge data aspects accordingly, and **(d)** choosing an appropriate data revision for a user's analysis task.

2.2. Discussion

We presented a VA approach that leverages data quality information over time to allow the inspection of the development of a dataset. Users thus are able to inspect different states of open datasets to judge if the quality of a particular revision is properly structured, of satisfactory quality, and in a usable state. The recorded data provenance combines different meaningful data aspects and is used in our visual interactive methods to aid the user with assessment data quality. We plan to further elaborate the framework, and extend its interactive exploration features. If effectively integrated into current open data portals, our framework enables users to choose between multiple revisions of the data and retain access to quality information without having to individually inspect dataset revisions. So far, we only support tabular data, since the available datasets on the open data platform mainly focus on csv-Files, but we plan on supporting different data structures in the future, with particular focus on multivariate time series data. By providing provenance from data transformations, community-driven data processing could be expedited and access to difficult datasets could be facilitated.

Acknowledgements This work was supported by the Austrian Science Fund (FWF), Project No. I 2850-N31, Lead AgencyVerfahren (DACH) "Visual Segmentation and Labeling of Multivariate Time Series (VISSECT)".

References

[ade17] ADEQUATe Open Data Storage | Adequate Open Data, Feb. 2017. URL: https://www.adequate.at/adequate-open-data-storage. 1

[BKG*18] BORS C., KRIGLSTEIN S., GSCHWANDTNER T., MIKSCH S., POHL M.: Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics. *Journal of Data and Information Quality (JDIQ) Forthcoming* (Apr. 2018). 2

[GAM*14] GSCHWANDTNER T., AIGNER W., MIKSCH S., GÄRTNER J., KRIGLSTEIN S., POHL M., SUCHY N.: TimeCleanser: A Visual Analytics Approach for Data Cleansing of Time-oriented Data. In *Proc. of the 14th International Conference on Knowledge Technologies and Data-driven Business* (New York, NY, USA, 2014), i-KNOW '14, ACM, pp. 18:1–18:8. doi:10.1145/2637748.2638423. 1

[GGAM12] GSCHWANDTNER T., GÄRTNER J., AIGNER W., MIKSCH S.: A Taxonomy of Dirty Time-Oriented Data. In *Lecture Notes in Computer Science (LNCS 7465): Multidisciplinary Research and Practice for Information Systems (Proceedings of the CD-ARES 2012)* (Prague, Czech Republic, 2012), Quirchmayr G., Basl J., You I., Xu L., Weippl E., (Eds.), Springer, Berlin / Heidelberg, pp. 58–72. doi:10.1007/978-3-642-32498-7_5. 1

[KHP*11] KANDEL S., HEER J., PLAISANT C., KENNEDY J., VAN HAM F., RICHE N. H., WEAVER C., LEE B., BRODBECK D., BUONO P.: Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization Journal 10*, 4 (2011), 271–288. 1

[KPP*12] KANDEL S., PARIKH R., PAEPCKE A., HELLERSTEIN J. M., HEER J.: Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (New York, NY, USA, 2012), AVI '12, ACM, pp. 547–554. doi:10.1145/2254556.2254659. 1

[MGM12] MORRIS T., GUIDRY T., MAGDINIER M.: OpenRefine, 2012. URL: http://openrefine.org. 2

[RESC16] RAGAN E. D., ENDERT A., SANYAL J., CHEN J.: Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 31–40. doi:10.1109/TVCG.2015.2467551. 1

[Sad13] SADIQ S. (Ed.): *Handbook of Data Quality*. Springer Verlag, Berlin, Heidelberg, 2013. 1